

Histogram comparison as a powerful tool for the search of new physics at LHC. Application to CMSSM

Maria Eugenia Cabrera

*Instituto de Física Teórica, IFT-UAM/CSIC
U.A.M., Cantoblanco,
28049 Madrid, Spain*

J. Alberto Casas

*Instituto de Física Teórica, IFT-UAM/CSIC
U.A.M., Cantoblanco,
28049 Madrid, Spain*

Vasiliki A. Mitsou

*Instituto de Física Corpuscular, IFIC-UV/CSIC
Valencia, Spain*

Roberto Ruiz de Austri

*Instituto de Física Corpuscular, IFIC-UV/CSIC
Valencia, Spain*

Juan Terrón

*Departamento de Física Teórica C-15,
Universidad Autónoma de Madrid,
Madrid, Spain*

ABSTRACT:

We propose a rigorous and effective way to compare experimental and theoretical histograms, incorporating the different sources of statistical and systematic uncertainties. This is a useful tool to extract as much information as possible from the comparison between experimental data with theoretical simulations, optimizing the chances of identifying New Physics at the LHC. We illustrate this by showing how a search in the CMSSM parameter space, using Bayesian techniques, can effectively find the correct values of the CMSSM parameters by comparing histograms of events with multijets + missing transverse momentum displayed in the effective-mass variable. The procedure is in fact very efficient to identify the true supersymmetric model, in the case supersymmetry is really there and accessible to the LHC.

KEYWORDS: Beyond Standard Model, Supersymmetry Phenomenology, histogram comparison, Bayesian statistics.

Contents

1. Introduction	1
2. Comparison of histograms. Statistical uncertainties	3
2.1 Basic Ingredients and Notation	3
2.2 Computation of the likelihood	4
2.3 Separation of normalization and shape tests	5
3. Incorporating other sources of uncertainty. Systematic errors	6
3.1 General strategy	6
3.2 Ansätze for the transfer functions	7
3.3 Separation of normalization and shape tests	9
4. Application to the CMSSM	10
4.1 Set up	10
4.2 Results	15
5. Conclusions	21
6. Appendix A: Histogram comparison when experiment and the simulation have different luminosities	23

1. Introduction

The LHC is already probing new physics beyond the reach of past experiments. At any stage of this enterprise, i.e. with the available data at any time, there are two main questions to address: 1) Is there any signal of New Physics (NP)? and 2) In the positive case, which NP is it? In order to optimize the answer to these questions there is an intense activity to explore assorted strategies for the search of NP. The task is challenging, due in part to the fact that LHC data, though very rich, are not as clean as those from an e^+e^- collider. Besides, the theoretical calculations are also subject to great uncertainties and rely to some extent on Monte Carlo simulations.

Most of the LHC data can be organized in form of histograms with number of events of a certain kind (e.g. those presenting multijets + missing transverse momentum) displayed in different variables [1, 2, 3]: M_{eff} , p_T^{miss} , α_T , etc. In many cases the comparison with the simulations is done just by comparing the total number of events after performing different cuts in the variables involved. In this way, both ATLAS and CMS have already posed meaningful bounds [4, 5] on NP scenarios, in particular on the simplest supersymmetric model, the so-called Constrained Minimal Supersymmetric Standard Model (CMSSM) [6]. More precisely, the most powerful bounds on the CMSSM have been obtained by considering events with several jets + missing transverse momentum. Somehow, the study of the total number events, choosing different cuts on the sets of data, amounts to partially compare the shapes of the experimental and the theoretical (simulated) histograms; although in a way which is not optimal.

As mentioned above, even if we are quite sure to have a signal of NP, we face the problem of identifying the model producing such signal. Of course the variety of scenarios of NP is enormous, which makes the job very complex. Even playing in the framework of a given scenario, such as the CMSSM, the sole study of the number of events of a certain kind is not enough to determine the parameters of the model, due to the existence of big degeneracies in such determination. Again, this situation can be improved by probing different cuts in the sets of data. But, once more, this is not an optimized way of comparing theory and experiment, since the richness of the data is not completely exploited.

The goal of this paper is to propose an effective and rigorous way to compare experimental and theoretical histograms, incorporating the different sources of uncertainty involved in the task. In our opinion, in an experiment with the characteristics of the LHC this is a useful tool to extract as much information as possible from the comparison between experimental data with theoretical simulations. We illustrate this usefulness by showing how a search in the CMSSM parameter space, using Bayesian techniques, can effectively find the correct values of the CMSSM parameters by comparing histograms of events with multijets + missing transverse momentum displayed in the M_{eff} variable. This procedure could be very efficient to identify the true supersymmetric model, in the case supersymmetry is really there and accessible to the LHC.

In section 2 we establish the notation and the statistical basis for the rigorous comparison between the experimental and the theoretical histograms. Section 3 is devoted to the incorporation of extra sources of uncertainty, in particular systematic ones. At the end of this section we give our final formula for the complete likelihood of a theoretical model by histogram comparison. In section 4 we illustrate the proposed technique by showing how a search in the constrained-MSSM parameter space, using Bayesian techniques, can effectively find the correct values of the MSSM parameters by comparing histograms of events with multijets + missing transverse momentum displayed in the effective-mass variable. But, of course, the technique can be applied to any scenario of new physics. Our conclusions are summarized in section 5. Finally, in the Appendix we show how our final formula for histogram-comparison is (slightly) modified when the effective luminosity of the theoretical simulation is not the same as the experimental one.

2. Comparison of histograms. Statistical uncertainties

2.1 Basic Ingredients and Notation

Suppose we have experimental data, e.g. multijet + missing transverse momentum events at LHC, organized in an histogram upon some variable M , e.g. the effective mass of the events, as defined in ref. [2]. Let us call K the number of bins of the histogram. Each bin corresponds to a (central) value of the effective mass, M_i . We will denote the bin contents (number of events for each M_i) by v_i . The total number of events is $v = \sum_i v_i$.

Leaving apart for the moment all sources of systematic uncertainties, the probability that the experiment produces the actual data, v_i , is given by a Poisson distribution

$$\mathcal{P}(v_i) = \prod_{i=1}^K \frac{\nu_i^{v_i}}{v_i!} e^{-\nu_i}, \quad (2.1)$$

where ν_i are the expected values (or “means”) of the distribution. The values of ν_i are in principle calculable (at some degree of precision) provided we knew the theory responsible for them, e.g. the Standard Model. But we are precisely trying to uncover unknown NP, therefore ν_i are unknown.

On the other hand, working within a scenario of NP defined by some parameters, θ_a (for example the parameters of the CMSSM), we can in principle calculate the means under, supposedly, the same conditions of energy and luminosity as the experiment. We will denote μ_i these theoretical means. Of course, μ_i depend on the point in the parameter space, i.e. the precise model under consideration. If the model is the true one, then $\nu_i = \mu_i$. This is the so-called “null-hypothesis”. The likelihood of a point of the parameter space is the corresponding probability of producing the observed data, v_i , under the null-hypothesis, i.e.

$$\mathcal{P}(v_i) = \prod_{i=1}^K \frac{\mu_i^{v_i}}{v_i!} e^{-\mu_i}. \quad (2.2)$$

The likelihood is a crucial quantity to compare the viability of the different regions of the parameter space, both in frequentist and Bayesian analyses (see, e.g. [7]). In particular, in Bayesian analyses one is interested in determining the probability density of a point of the parameter space, θ_a , given an experimental set of data (in our case, v_i). This is the so-called posterior probability density function (pdf), $p(\theta_a|\text{data})$, which is given by the fundamental Bayesian relation

$$p(\theta_a|\text{data}) = p(\text{data}|\theta_a) p(\theta_a) \frac{1}{p(\text{data})}. \quad (2.3)$$

Here $p(\text{data}|\theta_a)$ is the above-mentioned likelihood, i.e. the probability of obtaining the observed data if the model defined by the θ_a parameters is the true one; while $p(\theta_a)$ is the prior, i.e. the “theoretical” probability density that we assign a priori to the point

in the parameter space; and $p(\text{data})$ is a normalization factor that ensures that the total probability is one.

In order to compute the likelihood (2.2) we need the theoretical means, μ_i . However, in practice one does not have at disposal a complete evaluation of μ_i , but rather a simulation of the process using diverse computation codes. The results of the simulation can also be organized in an histogram with K bins, associated with the same values of the effective mass, M_i . The bin contents of the simulation are denoted by u_i , with total number of events $u = \sum_i u_i$. Of course, the values of u_i obey also a Poisson statistics

$$\mathcal{P}(u_i) = \prod_{i=1}^K \frac{\mu_i^{u_i}}{u_i!} e^{-\mu_i}. \quad (2.4)$$

Here we have again left aside for the moment all sources of systematic uncertainties associated with the theoretical simulation.

2.2 Computation of the likelihood

As mentioned, usually the codes provide values for u_i , but not for μ_i . If we had enough computation time we could obtain a good evaluation of the theoretical means, μ_i , since, increasing the statistics, the bin contents would approach the mean values with decreasing relative uncertainty. This would be practical if we knew from the beginning which specific model we want to test, but this procedure is not efficient if we want to scan the parameter space, testing thousands or millions of models (points in that space). So, identifying u_i with μ_i is not justified unless u_i is large. The relation between them is given by eq.(2.4). Since we are not sure about the values of μ_i , we cannot directly calculate the likelihood $\mathcal{P}(v_i)$ from eq.(2.2). The best we can do is calculate $\mathcal{P}(v_i|u_i)$, i.e. the probability of getting the experimental data, v_i , under the assumption that the model is the true one (null-hypothesis), given that the simulation has produced u_i ,

$$P(v_i|u_i) = \int \prod_{i=1}^K d\mu_i \mathcal{P}(v_i|\mu_i) \mathcal{P}(\mu_i|u_i). \quad (2.5)$$

Here $\mathcal{P}(v_i|\mu_i)$ is given by the Poisson distribution (2.2) and $\mathcal{P}(\mu_i|u_i)$ denotes the probability that the theoretical means are μ_i , given that the simulation has produced the u_i -histogram. $\mathcal{P}(\mu_i|u_i)$ is not known, we must infer it using the Bayes theorem,

$$P(\mu_i|u_i) = \frac{\mathcal{P}(u_i|\mu_i)\mathcal{P}(\mu_i)}{\int d\mu_i \mathcal{P}(u_i|\mu_i)\mathcal{P}(\mu_i)}, \quad (2.6)$$

where $\mathcal{P}(u_i|\mu_i)$ is the probability for each individual bin, given by the Poisson distribution (2.4), and $\mathcal{P}(\mu_i)$ is the prior for μ_i . Since $\mathcal{P}(u_i|\mu_i)$ is peaked around $u_i = \mu_i$, the dependence on the prior, $\mathcal{P}(\mu_i)$, is small, but nevertheless it is there. The simplest procedure here is to take a flat prior for $\mathcal{P}(\mu_i)$. Then the $\mathcal{P}(\mu_i)$ cancels in the numerator and the denominator of eq.(2.6) (the latter becomes simply 1), and we can identify

$$\mathcal{P}(\mu_i|u_i) \equiv \mathcal{P}(u_i|\mu_i). \quad (2.7)$$

Now eq.(2.5) reads

$$P(v_i|u_i) = \int \prod_{i=1}^K d\mu_i \frac{\mu_i^{v_i}}{v_i!} e^{-\mu_i} \frac{\mu_i^{u_i}}{u_i!} e^{-\mu_i} = \prod_{i=1}^K \frac{(u_i + v_i)!}{u_i! v_i!} 2^{-1-u_i-v_i}. \quad (2.8)$$

This formula represents our best estimate the likelihood, although is only valid when the non-statistical sources of uncertainty, both in the experimental and in the theoretical side, are ignored (they are incorporated in the next section). Note that expression (2.8) avoids the problem of the empty bins in the theoretical simulation. In other words, if one simply identified $\mu_i = u_i$, then the presence of an empty bin ($u_i = 0$) would make the whole likelihood - eq.(2.2)- vanishing. Therefore the $\mathcal{P}(\mu_i|u_i)$ piece in the calculation of the likelihood, eq.(2.5), is important, at least for bins with low statistics.

2.3 Separation of normalization and shape tests

Suppose for a moment we could calculate all μ_i with great accuracy and that we keep ignoring other sources of uncertainties different from the statistical ones. Then, the likelihood is simply given by the Poisson distribution $\mathcal{P}(v_i)$, as given by eq.(2.2).

Now, it is interesting that that expression can be separated in a test for the global normalization (the total number of events) and a test for the shape. Namely

$$P(v_i) = \prod_{i=1}^K \frac{\mu_i^{v_i}}{v_i!} e^{-\mu_i} = \mathcal{P}(\text{norm}) \times \mathcal{P}(\text{shape}), \quad (2.9)$$

where

$$\begin{aligned} \mathcal{P}(\text{norm}) &= \frac{\mu^v}{v!} e^{-\mu} \quad \text{with } \mu = \sum_i \mu_i, \\ \mathcal{P}(\text{shape}) &= \prod_{i=1}^K \frac{\mu_i^{v_i}}{\mu^{v_i}} \frac{(v!)^{1/K}}{v_i!} = V \prod_{i=1}^K \left(\mu_i \frac{v}{\mu} \right)^{v_i}, \end{aligned} \quad (2.10)$$

with

$$V = \frac{v!}{v^v} \prod_{i=1}^K \frac{1}{v_i!} = \text{const.} \quad (2.11)$$

Notice that both $\mathcal{P}(\text{norm})$, $\mathcal{P}(\text{shape})$ are proportional to terms given by Poisson distributions. In particular, $\mathcal{P}(\text{shape})$ is proportional to a Poisson distribution, where the means μ_i are re-normalized so that they would fit perfectly the total number of events:

$$\begin{aligned} \mathcal{P}(\text{shape}) &= V \prod_{i=1}^K \left(\mu_i \frac{v}{\mu} \right)^{v_i} \\ &= (e^v \prod_{i=1}^K v_i!) V \prod_{i=1}^K \frac{(\mu_i \frac{v}{\mu})^{v_i}}{v_i!} e^{-\mu_i \frac{v}{\mu}} \\ &= \frac{v! e^v}{v^v} \prod_{i=1}^K \frac{(\mu_i \frac{v}{\mu})^{v_i}}{v_i!} e^{-\mu_i \frac{v}{\mu}}. \end{aligned} \quad (2.12)$$

This expression is really independent of the global normalization. i.e. if we make $\mu_i \rightarrow a\mu_i$, then $\mathcal{P}(\text{shape})$ remains the same. This also tells us that if we fix the shape of a simulated histogram and allow to change its global normalization (i.e. we allow $\mu_i \rightarrow a\mu_i$), the total probability (2.9) is always maximal when the global mean, μ , coincides with the total number of events, v .

The interesting thing about separating normalization and shape tests is that one can treat the extra sources of systematic uncertainty for both in a separate way, as will become clear in the next section. For instance, one may consider that the amount of systematic uncertainty in the global normalization is larger than in the shape, and hence it is useful to separate the two tests.

3. Incorporating other sources of uncertainty. Systematic errors

3.1 General strategy

There are several sources of uncertainty in the comparison of the experimental data with the theoretical predictions. First, there is the statistical uncertainty, associated to the Poisson distributions, which has been the subject of the previous section. Besides there are additional sources of systematic uncertainty, both in the experimental side (the resolution and the scale of jets and missing transverse momentum, b-tagging, pile-up, etc.) and in the theoretical one (K-factors, parton distribution functions, etc.). However, for practical purposes, we can treat the experimental data as if they were free from systematic errors and “absorb” *all* the experimental systematic uncertainty in the theoretical side.

We will call μ_i^{th} the means that, in the simulation process, have produced the theoretical (u_i) histogram. Now, due to the systematic uncertainty, we *cannot* identify them directly with the “true means”, μ_i , which are the real ones associated with the model under consideration, and thus the ones that, supposedly, have “produced” the experimental histogram (v_i) under the null-hypothesis. The relation between them can be expressed as

$$\mu_i(M) = F(M_i) \mu_i^{th}, \quad (3.1)$$

where $F(M)$ is some “transfer function” on the effective mass (M) that encodes all (experimental and theoretical) systematic uncertainties. This function can depend on a number of unknown parameters, though we know it cannot be completely arbitrary (below we give an ansatz for $F(M)$).

Now, in analogy with eq.(2.5), the best estimate for the likelihood is

$$\mathcal{P}(v_i|u_i) = \int DF \int D\mu_i^{th} \mathcal{P}(v_i|\mu_i) \mathcal{P}(\mu_i^{th}|u_i) \mathcal{P}(\mu_i|\mu_i^{th}), \quad (3.2)$$

where $\mathcal{P}(\mu_i|\mu_i^{th}) \equiv \mathcal{P}(F)$ is still to be guessed, and the integration measure DF is written in a symbolic form. The first two factors in the integrand are statistical probabilities, as in (2.5). The third factor contains the systematic uncertainty (if we decided to ignore it, then we would simply take $\mathcal{P}(\mu_i|\mu_i^{th}) \equiv \delta(F - 1)$).

We can write explicit expressions for the three factors in eq.(3.2). The first factor, $\mathcal{P}(v_i|\mu_i)$, is given by the Poisson distribution (2.2). Regarding the second factor, recall that (taking a flat prior for μ_i^{th}) we can identify

$$\mathcal{P}(\mu_i^{th}|u_i) \sim \mathcal{P}(u_i|\mu_i^{th}) = \prod_{i=1}^K \frac{(\mu_i^{th})^{u_i}}{u_i!} e^{-\mu_i^{th}}. \quad (3.3)$$

Finally, we have to make ansatz for the F function and its probability, $\mathcal{P}(\mu_i|\mu_i^{th}) \equiv \mathcal{P}(F)$. Since it is convenient to separate the uncertainties associated to the global normalization and to the shape, we express eq.(3.1) as

$$\mu_i = F(M_i) \mu_i^{th} = f g_i \mu_i^{th}. \quad (3.4)$$

Here f and g_i carry the uncertainty in the global normalization and in the shape, respectively. With this definition, g_i obey the relation

$$\sum_i g_i \mu_i^{th} = \sum_i \mu_i^{th} \equiv \mu^{th}, \quad (3.5)$$

i.e. the g_i parametrize systematic errors that modify the shape of the histogram without changing the total number of events. The situation $f = g_i = 1$ corresponds to the absence of systematic errors, but we have to assign a non-vanishing probability to the possibility that f, g_i depart from that ideal situation. Thus we write

$$\mathcal{P}(\mu_i|\mu_i^{th}) \equiv \mathcal{P}(f, g_i) = \mathcal{P}(f) \mathcal{P}(g). \quad (3.6)$$

For the moment we do not write a concrete ansatz for $\mathcal{P}(f)$, $\mathcal{P}(g)$ (this is postponed to the next subsection). So, the likelihood (3.2) is given by

$$\mathcal{P}(v_i|u_i) = \int D\mu_i^{th} \int Df Dg \left(\prod_{i=1}^K \frac{\mu_i^{v_i}}{v_i!} e^{-\mu_i} \right) \left(\prod_{i=1}^K \frac{(\mu_i^{th})^{u_i}}{u_i!} e^{-\mu_i^{th}} \right) \mathcal{P}(f) \mathcal{P}(g), \quad (3.7)$$

In this expression Df , Dg are symbolic ways to express integration over all the possibilities for f , g_i .

3.2 Ansätze for the transfer functions

In eq.(3.4) we have written the “transfer” function, F , that encodes the systematic uncertainty, as

$$F(M_i) = f g_i, \quad (3.8)$$

but so far we have not established on which parameters the F -function –and thus the quantities f , g_i – depend. A simple and handy choice for practical purposes is to take the very values of $\{f, g_i\}$ as those independent parameters. Alternatively, since systematic errors must depend on M in a smooth way, we could parametrize $F(M)$ as a smooth function, e.g. $F \sim f \sum_{\alpha} a_{\alpha} P_{\alpha}$, where P_{α} are \sim Legendre Polynomials and the summation

contains just a few terms. Then, the F function would be defined by the a_α coefficients (together with the global normalization factor, f). This would be sensible, but it leads to very cumbersome expressions, difficult to handle. On the other hand, since in practice v_i and u_i are both quite smooth (apart from statistical noise), only sets of values of $F(M_i)$ that vary smoothly with M can lead to a simultaneous fit of both histograms. In other words, chaotic values of $F(M_i)$ (or, equivalently, g_i) will be strongly penalised by the $\mathcal{P}(v_i|\mu_i)$ piece (first factor in eq.(3.7)). So, even if those eccentric choices for g_i are not specially penalised by $\mathcal{P}(g)$, they are by other factors in the likelihood and become irrelevant. In consequence, choosing $\{f, g_i\}$ as independent parameters is a reasonable option

Concerning the integration measures, we could simply take $Df = df$, $Dg = \prod_i dg_i$. However, since $\{f, g_i\}$ are defined as multiplicative factors in eq.(3.4), it seems much more sensible to use their magnitudes as the actual unknowns. This is equivalent to choose $\{\ln f, \ln g_i\}$ as the independent parameters. Then,

$$Df \equiv \frac{1}{f} df, \quad Dg \equiv \prod_{i=1}^K \frac{1}{g_i} dg_i. \quad (3.9)$$

Of course, since $\{f, g_i\}$ are never far from 1, it does not make a big difference to use $\{f, g_i\}$ or $\{\ln f, \ln g_i\}$, but it can be checked that the second option leads to a more stable and satisfactory test. Note that, in principle, the g_i variables are subject to condition (3.5), so there are in fact $K - 1$ independent g_i variables. However, for the moment we have ignored such complication in writing (3.9).

Finally, concerning the probabilities $\mathcal{P}(f)$, $\mathcal{P}(g)$, we can take them as gaussians centered around $f = g_i = 1$. The argument of these gaussians must be essentially the “squared-distance” of $\{f, g_i\}$ to their central values, i.e. $\mathcal{P}(f) \sim \exp\{-\frac{1}{2}(f - 1)^2\}$ and $\mathcal{P}(g) \sim \exp\{-\frac{1}{2} \int dM (g(M) - 1)^2\} \sim \exp\{-\frac{1}{2} \sum_i (g_i - 1)^2\}$. A nice fact here is that $\mathcal{P}(g)$ appears naturally factorized as $\prod_i \mathcal{P}(g_i)$, which is very convenient for analytical manipulations.

A suitable (and equivalent at first order), way to express these ansätze is by using the logarithmic variables, $\{\ln f, \ln g_i\}$:

$$\mathcal{P}(f) = \frac{1}{\sqrt{2\pi}\Delta_f} e^{-\frac{1}{2}\left(\frac{\ln f}{\Delta_f}\right)^2}, \quad (3.10)$$

$$\mathcal{P}(g) \propto \frac{1}{\Delta_g^K} e^{-\frac{1}{2}\sum_i \left(\frac{\ln g_i}{\Delta_g}\right)^2}, \quad (3.11)$$

where the widths Δ_f , Δ_g measure our degree of ignorance about the magnitude of f , g_i . Note that the use of logarithmic variables allows to maintain the whole range of integration of the gaussians, $[-\infty, \infty]$, without artificial cuts to keep $\{f, g_i\}$ positive.

In any case, we will go as far as possible in the analysis without specifying the precise ansätze for $\mathcal{P}(f)$, $\mathcal{P}(g)$.

3.3 Separation of normalization and shape tests

Coming back to our expression (3.7) for the likelihood, we note that the first factor of (3.7) may be decomposed, as in eqs.(2.9)-(2.10), into a factor for the global-normalization and another for the shape:

$$\left(\prod_{i=1}^K \frac{\mu_i^{v_i}}{v_i!} e^{-\mu_i} \right) = \frac{\mu^v}{v!} e^{-\mu} V \prod_{i=1}^K \left(\mu_i \frac{v}{\mu} \right)^{v_i}, \quad (3.12)$$

where V is given in eq.(2.11). Since the *total* number of events is normally large the global-normalization factor can be approximated by a Dirac delta,

$$\frac{\mu^v}{v!} e^{-\mu} \simeq \delta(\mu - v) = \delta(f\mu^{th} - v) = \frac{1}{\mu^{th}} \delta(f - v/\mu^{th}). \quad (3.13)$$

Analogously, the second factor of (3.7) can be written as

$$\prod_{i=1}^K \frac{(\mu_i^{th})^{u_i}}{u_i!} e^{-\mu_i^{th}} \simeq \delta(\mu^{th} - u) \times U \prod_{i=1}^K \left(\mu_i^{th} \frac{u}{\mu^{th}} \right)^{u_i}, \quad (3.14)$$

with

$$U = \frac{u!}{u^u} \prod_{i=1}^K \frac{1}{u_i!} = \text{const.} \quad (3.15)$$

We can use the presence of these deltas to extract pieces of the integrand of eq.(3.7) outside the sign of integration. Hence,

$$\begin{aligned} \mathcal{P}(v_i|u_i) &\propto \mathcal{P}(f = \frac{v}{u}) \\ &\times \int D\mu_i^{th} Dg \left(\prod_{i=1}^K \frac{(\frac{v}{u} g_i \mu_i^{th})^{v_i}}{v_i!} e^{-\frac{v}{u} g_i \mu_i^{th}} \right) \left(\prod_{i=1}^K \frac{(\mu_i^{th})^{u_i}}{u_i!} e^{-\mu_i^{th}} \right) \mathcal{P}(g). \end{aligned} \quad (3.16)$$

Note that we have made explicetly the integration in $\int Df = \int (1/f) df$, but not in $\int d\mu^{th}$. However the implicit presence of the $\delta(\mu^{th} - u)$ in the integrand, as expressed in eq.(3.14), has allowed us to replace $\mu^{th} \rightarrow u$ in a consistent way.

Assuming in the previous expression that Dg and $\mathcal{P}(g)$ are factorizable as products of K factors, like in eqs.(3.9), (3.11), makes much easier the integration in practice. As mentioned, g_i are subject to condition (3.5), so strictly speaking we only have $K - 1$ independent g_i variables and this factorization is not complete. In spite of this, assuming a complete factorization is a sensible and good approximation. The reason is the following. In eq.(3.16) the Poisson distribution in the first factor of the integrand only departs appreciably from zero when $\sum g_i \mu_i^{th} \simeq u \simeq \mu^{th}$. This can be checked by doing again a decomposition of such distribution as in (2.9, 2.10) and noting that the global-normalization piece of the decomposition is essentially a Dirac delta, $\delta(\sum \frac{v}{u} g_i \mu_i^{th} - v)$. So, even assuming that Dg and $\mathcal{P}(g)$ are factorizable, and thus integrating over sets of $\{g_1, g_2, \dots, g_K\}$ which do

not respect (3.5), the Poisson distribution in the first factor of the integrand causes that only those sets that obey condition (3.5) will contribute appreciably to the integral. Alternatively one could understand this procedure considering that in the expression (3.4), the g_i variables that encode M_i -dependent systematic errors can also distort the total number of events. This is in fact a quite realistic situation. Then, the relation (3.5) is not to be imposed and Dg and $\mathcal{P}(g)$ become factorizable as products of K factors (strictly speaking). The trouble is that the previous separation between normalization and shape cannot be done exactly. But, if the g_i only amount to slight distortions of the total normalization (in other words, $\mathcal{P}(g)$ penalizes much more severely the variation in the normalization than $\mathcal{P}(f)$) the separation (3.14) is a good approximation and (3.16) is valid.

Now, taking profit of the factorization of Dg and $\mathcal{P}(g)$ we can make explicitly the integration in the μ_i^{th} variables, with no need of specifying the ansätze for $\mathcal{P}(f)$, $\mathcal{P}(g)$:

$$\begin{aligned} \mathcal{P}(v_i|u_i) &\propto \mathcal{P}(f = \frac{v}{u}) \\ &\times \prod_{i=1}^K \left(\frac{(u_i + v_i)!}{u_i!v_i!} \int dg_i \frac{1}{g_i} \left(\frac{v}{u} g_i \right)^{v_i} \left(1 + \frac{v}{u} g_i \right)^{-1-u_i-v_i} \mathcal{P}(g_i) \right). \end{aligned} \quad (3.17)$$

In this expression, the factor of the first line, $\mathcal{P}(f = \frac{v}{u})$, carries the test for the global normalization: it is only sensitive to the mismatch between the experimental total number of events, v , and the theoretical one, u . The remaining factor (second line) corresponds to the test of the shape. It is interesting to check that indeed, for given u, v , this expression has a maximum at $u_i = (u/v)v_i$.

Eq.(3.17) represents our final expression to evaluate the likelihood of a simulated histogram, u_i , confronted to the experimental one, v_i . (A modified version is given in eq.(6.4) of Appendix A to incorporate the fact that the luminosity of the simulated histogram may be different from that of the experimental one.) This expression amounts to realize K integrals, which can be done numerically at low cost in computing time, even if one needs to probe thousands or millions of histograms, corresponding to points in the parameter space of a theoretical scenario. All this is illustrated in the next section.

4. Application to the CMSSM

4.1 Set up

In this section we apply the previous histogram-comparison techniques to the study of the Minimal Supersymmetric Standard Model (MSSM) [13]. More precisely, we will consider the somehow standard framework, often called CMSSM or MSUGRA, in which the soft parameters are assumed universal at a high scale (M_X), where the supersymmetry (SUSY) breaking is transmitted to the observable sector; as happens e.g. in the gravity-mediated SUSY breaking scenario. Hence, our parameter-space is defined by the following parameters:

$$\{\theta_i\} = \{m, M_{1/2}, A, B, \mu, s\}. \quad (4.1)$$

Here m , $M_{1/2}$ and A are the universal scalar mass, gaugino mass and trilinear scalar coupling; B is the bilinear scalar coupling; μ is the usual Higgs mass term in the superpotential; and s stands for the SM-like parameters of the MSSM. The latter include the $SU(3) \times SU(2) \times U(1)_Y$ gauge couplings, g_3, g, g' , and the Yukawa couplings, which in turn determine the fermion masses and mixing angles. All the initial parameters are defined at the M_X scale.

The goal is to scan the CMSSM parameter space, determining the most probable region in it, given the available (present or future) experimental (mainly LHC) data. To show the power of the histogram-comparison technique, we will simulate LHC data assuming that nature lives in a standard benchmark SUSY model. This simulation will be considered as our (mock) experimental data. Then we will scan the CMSSM parameter space using Bayesian techniques to find out the most probable region of parameters, showing to which extent the histogram-comparison between the mock data and the theoretical prediction is capable to determine the “true” model.

As mentioned in sect. 2, in a Bayesian analysis the most important quantity is the *posterior* probability density function (pdf) in the parameter space, which is given by the fundamental Bayes’ relation

$$p(s, m, M_{1/2}, A, B, \mu | \text{data}) \propto p(\text{data} | s, m, M_{1/2}, A, B, \mu) p(s, m, M_{1/2}, A, B, \mu), \quad (4.2)$$

where the first factor in the right hand side is the *likelihood* (probability of measuring the observed data assuming that the point in the parameter space is the true model) and the second one is the *prior* (probability assigned to that point before knowing the experimental data). Next, we discuss the precise form of the pdf (4.2) for the problem at hand.

First of all, it should be noticed that very often in statistical problems not all the parameters that define the system are of the same interest. The usual technique to eliminate the less interesting ones from the problem is simply marginalizing them, i.e. integrating the pdf (4.2) in those variables (for a review see ref. [9]). This is the standard procedure to deal with the *nuisance parameters* $\{s\}$. Besides, for the purposes of scanning the CMSSM parameter space it is convenient to trade some of the initial parameters (4.1) by others with more direct phenomenological significance. We follow here the approach expounded in detail in refs. [10, 11]. As usual, the value of μ can be traded by the value of M_Z using the minimization conditions of the Higgs potential. The Yukawa couplings can be traded by the physical fermion masses; in particular the top Yukawa coupling, y_t , can be traded by the top mass, m_t . Finally, it is highly advantageous to trade the initial B -parameter by the derived $\tan \beta \equiv \langle H_u \rangle / \langle H_d \rangle$ parameter, where H_u, H_d are the Higgs doublets of the MSSM¹.

Consequently, to write the pdf in the new variables one should compute the Jacobian, J , of the transformation

$$\{\mu, y_t, B\} \rightarrow \{M_Z, m_t, \tan \beta\}. \quad (4.3)$$

¹This change of variables still leaves the sign of μ undetermined. For simplicity we have assumed a positive μ in the rest of the paper.

On the other hand, the μ -parameter (now traded by M_Z) can be easily marginalized, together with the nuisance parameters, taking profit of the high precision of our knowledge of M_Z^{exp} . Consequently, the final expression for the posterior (4.2) in the new variables is

$$p(m, M_{1/2}, A, \tan \beta | \text{data}) \propto J|_{\mu=\mu_Z} p(\text{data} | m, M_{1/2}, A, \tan \beta) \times p(m, M, A, B, \mu = \mu_Z), \quad (4.4)$$

where J is the Jacobian of the transformation (4.3) and μ_Z is the value of μ that reproduces M_Z^{exp} for the given values of $\{m, M, A, \tan \beta\}$. Let us discuss in order the three factors of the r.h.s. of (4.4).

The Jacobian factor, J , has to be evaluated using the (radiative) electroweak breaking conditions of the CMSSM. For the numerical analysis we have computed J using the **SoftSusy** code [12] which implements the full one-loop contributions and leading two-loop terms to the tadpoles for the those conditions, with parameters running at two-loops. This essentially corresponds to the next-to-leading log approximation. A quite accurate analytical expression of J , corresponding to the leading log approximation, reads [10]

$$J|_{\mu=\mu_Z} \propto \left[\frac{E}{R_\mu^2} \right] \frac{y}{y_{\text{low}}} \frac{\tan \beta^2 - 1}{\tan \beta (1 + \tan^2 \beta)} \frac{B_{\text{low}}}{\mu_Z}. \quad (4.5)$$

Here y denotes the top Yukawa coupling and the “low” subscript indicates that the quantity is evaluated at low scale (say $Q_{\text{low}} \equiv$ typical supersymmetric mass). R_μ and E are RG quantities, involved in the one-loop running of μ and y :

$$\mu_{\text{low}} = R_\mu \mu, \quad y_{\text{low}} \simeq \frac{y E(Q_{\text{low}})}{1 + 6y F(Q_{\text{low}})}, \quad (4.6)$$

where Q is the renormalization scale and $F = \int_{Q_{\text{high}}}^{Q_{\text{low}}} E \ln Q$. R_μ and E are definite functions that depend just on the top Yukawa coupling and the gauge couplings, respectively, [13, 14].

The important point about the Jacobian is that it does not represent any subjective prior on the parameters. Such subjectivity is still contained in the prior factor that stands in the second line of eq.(4.4). The Jacobian is simply a consequence of scanning the MSSM parameter space in some variables, which are not the initial ones, but derived quantities. Another important point is that J automatically incorporates a penalization of the regions that require fine-tuning in order to reproduce the correct electroweak scale (typically regions of large soft parameters), as well as a penalization of large $\tan \beta$, reflecting also the fine-tuning needed to implement such possibility. A more detailed discussion of these issues can be found in refs. [10, 11].

Let us discuss now the second factor of the r.h.s. of eq.(4.4), i.e. the likelihood. This consists of a product of likelihoods corresponding to the experimental observables used in the analysis. For the present one, we just consider (besides the value of M_Z^{exp}) the experimental bounds on the masses of supersymmetric particles and the lightest Higgs boson (see ref.[15] for details), and the mock LHC data of multijet events plus missing

transverse momentum, which is the main focus of this section². More precisely, for the sake of the simulated LHC data we work under the hypothesis that nature lies in the so-called SU9 benchmark point, defined in ref. [2]. This is specified by the following values of the CMSSM parameters:

$$m = 300 \text{ GeV}, \quad M_{1/2} = 425 \text{ GeV}, \quad A = 20 \text{ GeV}, \quad \tan \beta = 20, \quad \mu > 0. \quad (4.7)$$

The corresponding values of the squark mass (first two generations) and the gluino mass are $m_{\tilde{q}} = 920 \text{ GeV}$, $M_{\tilde{g}} = 994 \text{ GeV}$. This point of the CMSSM parameter space is on the verge of being excluded by the last analyses by ATLAS and CMS [16, 17]. Of course, assuming the SU9 point is just an example to show the histogram-comparison technique at work, combined with Bayesian analysis. The LHC simulation has been performed using Pythia version 6.419 [18] with events generated at $E_{CM} = 14 \text{ TeV}$, and selecting those satisfying the following cuts³:

- Three or more jets with $p_T > 30 \text{ GeV}$ and $|\eta| < 3.0$. The hardest with $p_T > 180 \text{ GeV}$ and $|\eta| < 1.7$, the second with $p_T > 110 \text{ GeV}$.
- $p_T^{\text{miss}} > 200 \text{ GeV}$.
- $\Delta\phi_1 > 0.3$, $\Delta\phi_2 > 0.3$, $\Delta\phi_3 > 0.3$.
- $\sqrt{\Delta\phi_2^2 + (\pi - \Delta\phi_1)^2} > 0.5$, $\sqrt{\Delta\phi_1^2 + (\pi - \Delta\phi_2)^2} > 0.5$ and $\Delta\phi_2 > \pi/9$.
- $H_T = \sum_{i=2} p_T^i + p_T^{\text{miss}} > 500 \text{ GeV}$.

where $\Delta\phi_i \equiv \Delta\phi(\text{jet}_i - p_T^{\text{miss}})$. Concerning the luminosity we have considered 10^4 supersymmetric events, upon which we impose the previous cuts. Since the total cross section for SUSY production in the SU9 model is 2.4 pb , this corresponds to a luminosity of about 4.2 fb^{-1} . The histogram of number of events as a function of the effective mass, M_{eff} , is shown in Fig. 1, where only the SUSY events have been displayed. For each event, M_{eff} is defined as [2]

$$M_{\text{eff}} = \sum_j |p_T^j| + p_T^{\text{miss}}. \quad (4.8)$$

where j runs over all jets satisfying the previous cuts. Note that the latter effectively imply a lower bound $M_{\text{eff}} \geq 680 \text{ GeV}$ for the events considered.

²Most of the electroweak precision tests are currently being surpassed by LHC data. Other observables, like $b \rightarrow s, \gamma$ would have a moderate impact in the analysis, but we want to focus on the impact of the LHC, which is already the dominant part of the likelihood. For the same reason we have not included the somewhat controversial $g - 2$ data or Dark Matter constraints.

³We have followed the strategy given in sect. 13.5 of ref. [1]

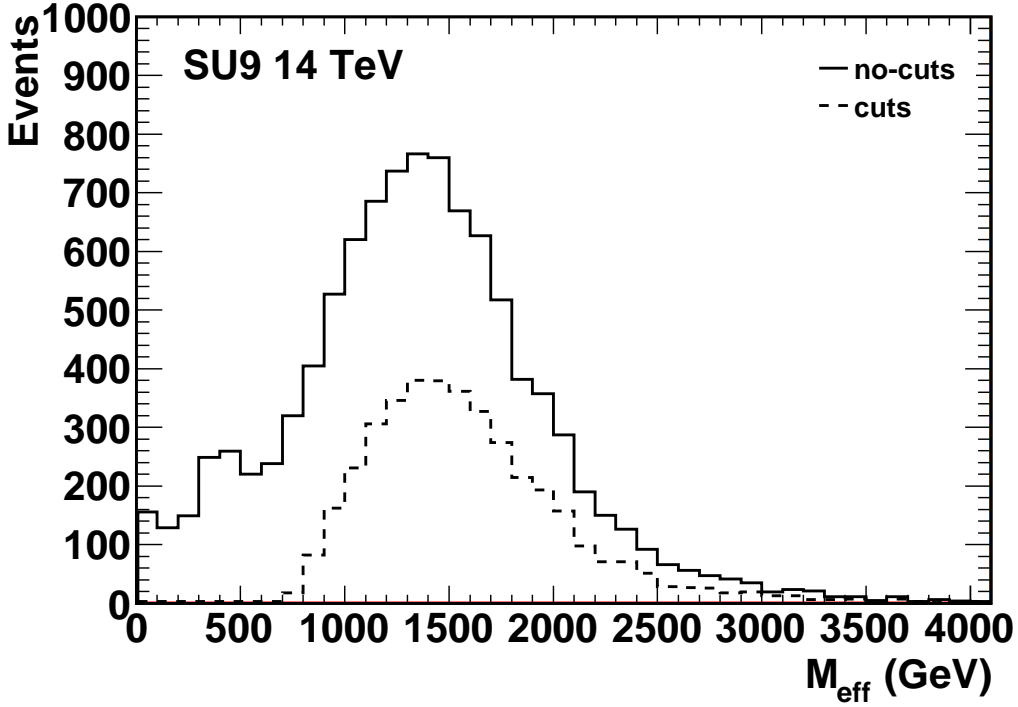


Figure 1: Effective mass distribution expected for the SUSY SU9 model without and with cuts applied as given in subsect. 4.1.

Using the notation of sects. 2, 3, the bin contents of the “experimental” histogram of Fig. 1 are the v_i quantities. For each point scanned in the parameter space we compute a simulated histogram (the u_i quantities in sects. 2, 3), and then evaluate the likelihood through eq. (3.17). In order to use that expression, we have to specify the $\mathcal{P}(f)$, $\mathcal{P}(g)$ functions, which encode the systematic uncertainty assigned to the total number of events and the shape of the histograms respectively. In our case, we have used for them the gaussian profiles (3.10), (3.11). The width of the first, Δ_f , reflects the uncertainty in the total number of events due to (mainly) theoretical uncertainties associated with the K-factors and the parton distribution functions. We have been pretty conservative, assuming $\Delta_f = 0.5$; in other words we accept that a factor 2 or 1/2 in the total number of events is plausible⁴. On the other hand, from eq.(3.11), Δ_g goes as the typical systematic uncertainty in the shape times \sqrt{K} . For instance if, *once* the uncertainty affecting the global normalization is extracted, one estimates that there remains a systematic uncertainty in the shape which is of order 10% for every bin, then one has to use $\Delta_g \simeq 0.1\sqrt{K}$ to reproduce that uncertainty at 1σ . In our case the total number of bins is $K = 10$, so we estimate $\Delta_g = 0.2$ as a reasonable choice.

⁴In the present scan we are using a tree-level Pythia simulation of the parton events; which implies an important uncertainty about the K-factors. In a 1-loop-refined simulation this uncertainty could be assumed smaller.

To conclude our discussion of the likelihood, let us note that the total cross section varies from point to point in the CMSSM parameter space. This implies that considering 10^4 initial supersymmetric events does not correspond to the same luminosity for each point scanned in the parameter space. But of course, the comparison of histograms must be realized under the same conditions of luminosity. This feature can be easily incorporated into the histogram-comparison technique discussed in sect. 3, leading to a slight and straightforward correction in the expression (3.17) for the likelihood, which now becomes eq.(6.4) of Appendix A (see that appendix for more details). And that is the expression that we have finally used to compute the likelihood when scanning the CMSSM parameter space.

We end up this subsection with a brief discussion of the third factor of the r.h.s. of eq.(4.4), i.e. the prior in the initial parameters. Admittedly, this is the less objective part of the statistical analysis, but one cannot simply ignore the prior. This would be equivalent to take a flat prior in the initial parameters, a choice which is as arbitrary as any other, unless one can give some argument of plausibility for it. Besides, in order to perform the marginalizations one should specify the ranges where the parameters live⁵. The dependence on the prior is actually a measure of the dependence of the results of the statistical analysis on a priori assumptions or prejudices. Note here that when the likelihood factor is very sharp, i.e. it distinctly selects a narrow region in the parameter space, the prior factor becomes irrelevant, since the posterior can only be sizeable where the likelihood is. But unfortunately we are not in that ideal situation yet, so there *exists* a dependence on the prior. Hence, the most conservative attitude is to use two different, though still reasonable, priors, and then compare the results. This gives a fair measure of the prior-dependence.

For that matter we have considered two somehow standard types of prior: flat and logarithmic. In a flat (logarithmic) prior one assumes that, in principle, the typical size (order of magnitude) of the soft terms can be anything, say from 10 GeV up to M_X , with equal probability. In our opinion, a logarithmic prior is probably the most reasonable option, since it amounts to consider all the possible magnitudes of the SUSY breaking in the observable sector on the same foot (this occurs e.g. in conventional SUSY breaking by gaugino condensation in a hidden sector). However, we will consider log and flat priors at the same level throughout the paper, in order to compare the results and thus evaluate the prior-dependence. The precise forms of these priors can be found in sect. 2 of ref. [11], together with a detailed discussion.

4.2 Results

We have computed the distribution of the posterior (4.4) in the CMSSM parameter space using a modified version of the public **SuperBayeS** package [19]⁶ adopting MultiNest v2.8 [20, 21] as a scanning algorithm. We use as running parameters a number of live points

⁵Fortunately, in our case this point is irrelevant thanks to the Jacobian factor, J . As mentioned above, J automatically incorporates a fine-tuning penalization of the high-energy region of the parameters, which thus becomes irrelevant. For more details see [11].

⁶For this paper, the public **SuperBayeS** code has been modified to interface with Pythia 6.419 [18].

$n_{\text{live}} = 2000$ and a tolerance parameter $\text{tol} = 1$. Our final inferences for each of the log and flat priors are obtained from chains generated with approximately 10^5 likelihood evaluations.

We have also included in our likelihood the limits on the lightest Higgs and SUSY masses provided by LEP⁷ and Tevatron . For details on the implementation see ref. [15].

For the marginalization procedure we have used $[0, M_X]$ as the range for m , $M_{1/2}$ and $|A|$. Besides, we have used $2 < \tan\beta < 62$. See a detailed discussion about this in sect. 2 of ref. [11].

In order to show the potential of the histogram-comparison technique, we have performed the analysis twice: switching off and on the shape test.

Test for the total number of events

First we compute the (LHC-part of the) likelihood associated with a particular point in the CMSSM parameter space by comparing the prediction for the total number of supersymmetric events, satisfying the cuts specified in the previous subsection, with the “experimental” result for that number (i.e. after subtracting the SM background).

This means that for each point examined we compute an histogram of events, u_i , but we just compare the total number, $u = \sum u_i$, with the experimental one, $v = \sum v_i$, using the first factor of eq.(3.17). More precisely, in order to incorporate the fact that the effective luminosity used in the simulation may change from point to point, we have actually used the –slightly modified– first factor of eq.(6.4),

$$\text{LHC – likelihood} \propto \mathcal{P}(f = \frac{v}{Lu}) , \quad (4.9)$$

where L is the quotient of the experimental luminosity and the luminosity of the simulation. We recall that the $\mathcal{P}(f)$ function carries all the uncertainties affecting the total number of events, except the purely statistical ones (which are subdominant when that number is large), and is given by the gaussian (3.10) with $\Delta_f = 0.5$; as explained in the previous subsection.

Fig. 2 (upper panels) shows the posterior pdf in the $M_{1/2}-m$ plane, after marginalizing the rest of the parameters: A , $\tan\beta$, together with the previously marginalized μ and the nuisance SM parameters, $\{s\}$. As discussed below, the cross section for the kind of events considered (multijets + missing transverse momentum) is actually fairly insensitive to the values of A , $\tan\beta$, so the marginalization in these parameters does not change appreciably the probability density in the $M_{1/2}-m$ plane. The left (right) panel corresponds to log (flat) priors for the soft terms. The shape of these plots can be easily understood. Since we are fitting a unique quantity, namely the total number of events, and we have two parameters, $\{M_{1/2}, m\}$, we can expect a degeneracy in the parameter space, which is in fact the case. The elongated shape of the allowed region, especially visible in the flat prior case, is in

⁷Recent LHC bound on the Higgs mass are still irrelevant to constrain the MSSM parameter space, though this situation will change soon [22]

fact a widening –due to the uncertainties– of the line where the degeneracy is exact, which includes of course the “true model”, i.e. SU9. This is marked with a red diamond in the plots. Note that for small gaugino mass, the squark masses become irrelevant, provided they are large enough, since in that case the dominant SUSY production are gluino pairs, whose masses do not depend on m . This is reflected in the vertical form of the region for small $M_{1/2}$.

Now, the shape of the line of degeneracy, somehow visible in the upper plots of Fig. 2, depends on the cuts used to select events. Note that, as the values of the soft terms get smaller, the cuts used to bound the energy of the first and second jets, and the missing transverse momentum (see previous subsection) become more and more inappropriate: many events with three or more jets plus missing transverse momentum do not pass the cuts. As a consequence the counted total number of events of this kind is dramatically cut out and can become equal to the experimental one. This enhances artificially the statistical weight of the low energy region. As a result the maximum value of the pdf, and its averaged central value (marked by a green dot), are shifted from the “true model” (marked by a red diamond).

There are ways to counteract these disagreeable effects. Playing with different cuts, the degeneracy gets partially broken and it is possible to discard larger regions of the parameter space. For instance, one can compare the total number of events using several choices for the lower bound on M_{eff} . Somehow, this equivaless to test the shape of the experimental and theoretical histograms, but not in the most efficient way. This is improved using the histogram-comparison technique explained in sections 2, 3, which we will apply shortly to this analysis.

Fig. 2 (lower panels) shows the posterior in the $\tan\beta - A$ plane, after marginalizing the rest of the parameters. As mentioned above, the cross section of the type of events considered does not depend appreciably on A and $\tan\beta$, and this is reflected in the plots. The preference for rather small values of both A and $\tan\beta$ is essentially a consequence of the Jacobian factor (4.5) in the posterior (4.4). As commented in the previous subsection the Jacobian automatically penalizes regions of the parameter space where fine-tuning is needed to reproduce the electroweak scale. This disfavors large values for both A and $\tan\beta$.⁸ The remarkable insensitivity to A and $\tan\beta$ is physically due to the fact that the CMSSM spectrum is not much dependent on the values of A and $\tan\beta$, except for mixing effects in the mass matrices of stops (and sbottoms and staus for large $\tan\beta$), charginos and neutralinos. Even for these matrices the effect is normally quite small. Thus the production rates of squarks and gluinos are quite independent of A and $\tan\beta$. Once the supersymmetric particles are created, their decay rates are not very relevant for the cross section of the process considered (multijets + missing transverse momentum), and, in any case, they are quite independent of these parameters too. This insensitivity to A and $\tan\beta$ could be partially cured by complementing the present analysis by a separate study of those events involving leptons [23], but that discussion is outside the scope of this paper.

⁸This is an statistical effect which is not visible in frequentist approaches, where the basic quantity is the likelihood and fine-tuning is not penalized, unless such penalization is artificially incorporated.

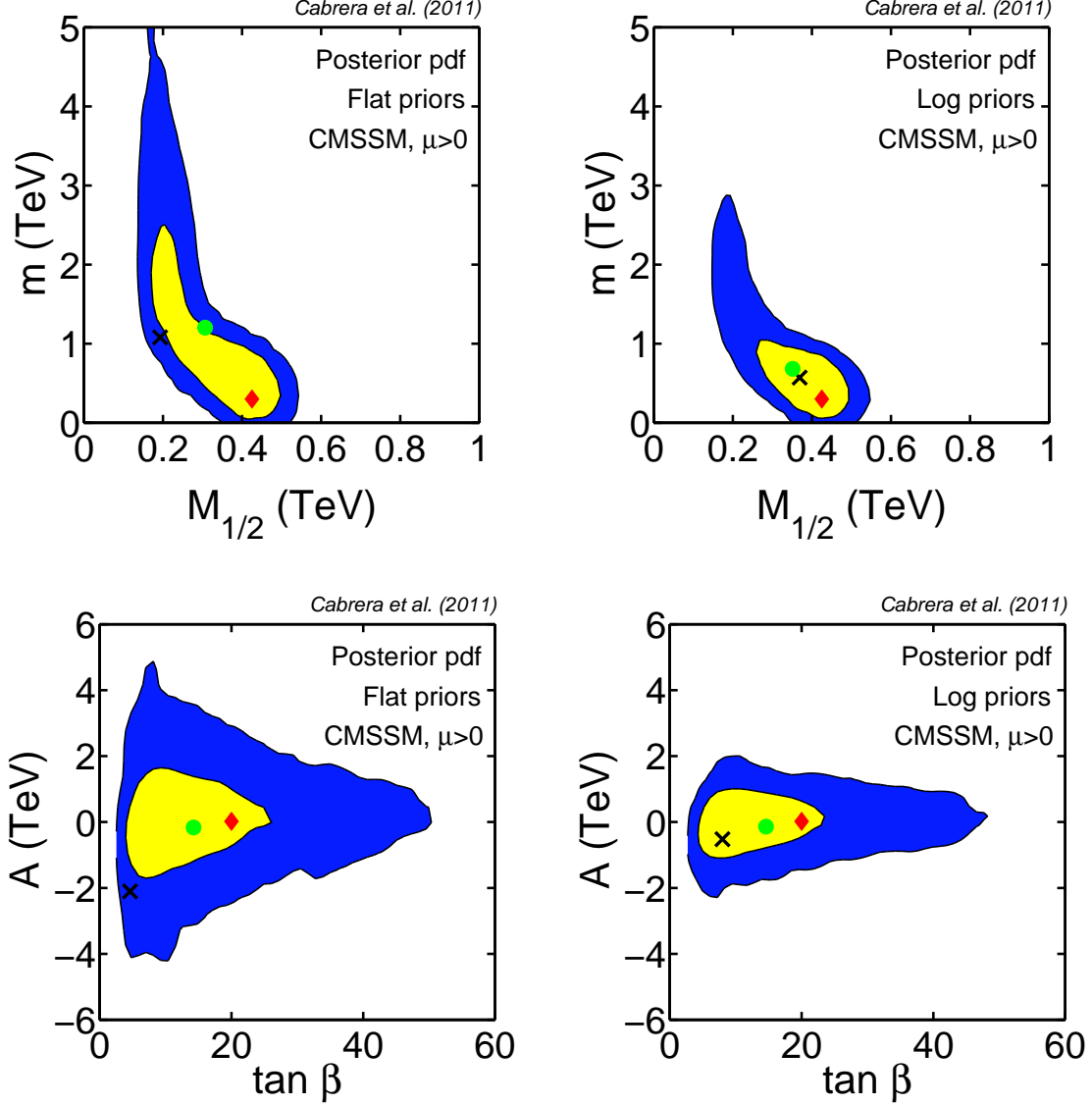


Figure 2: 2D marginalized posterior probability density function for flat (left panels) and logarithmic (right panels) priors using the normalization test. The inner and outer contours enclose respective 68% and 95% joint regions. The small filled circle represents the mean value of the posterior pdf, the cross corresponds to the best-fit point found and the diamond to the SU9 model, used to produce the mock experimental data.

Incorporation of the shape test

Now we repeat the analysis, but computing the likelihood associated with the LHC data with the use of the whole expression (3.17), which takes into account not only the total number of events, but also the comparison of the histogram shapes. Again, in order to incorporate the fact that the luminosity of the simulation changes from point to point in

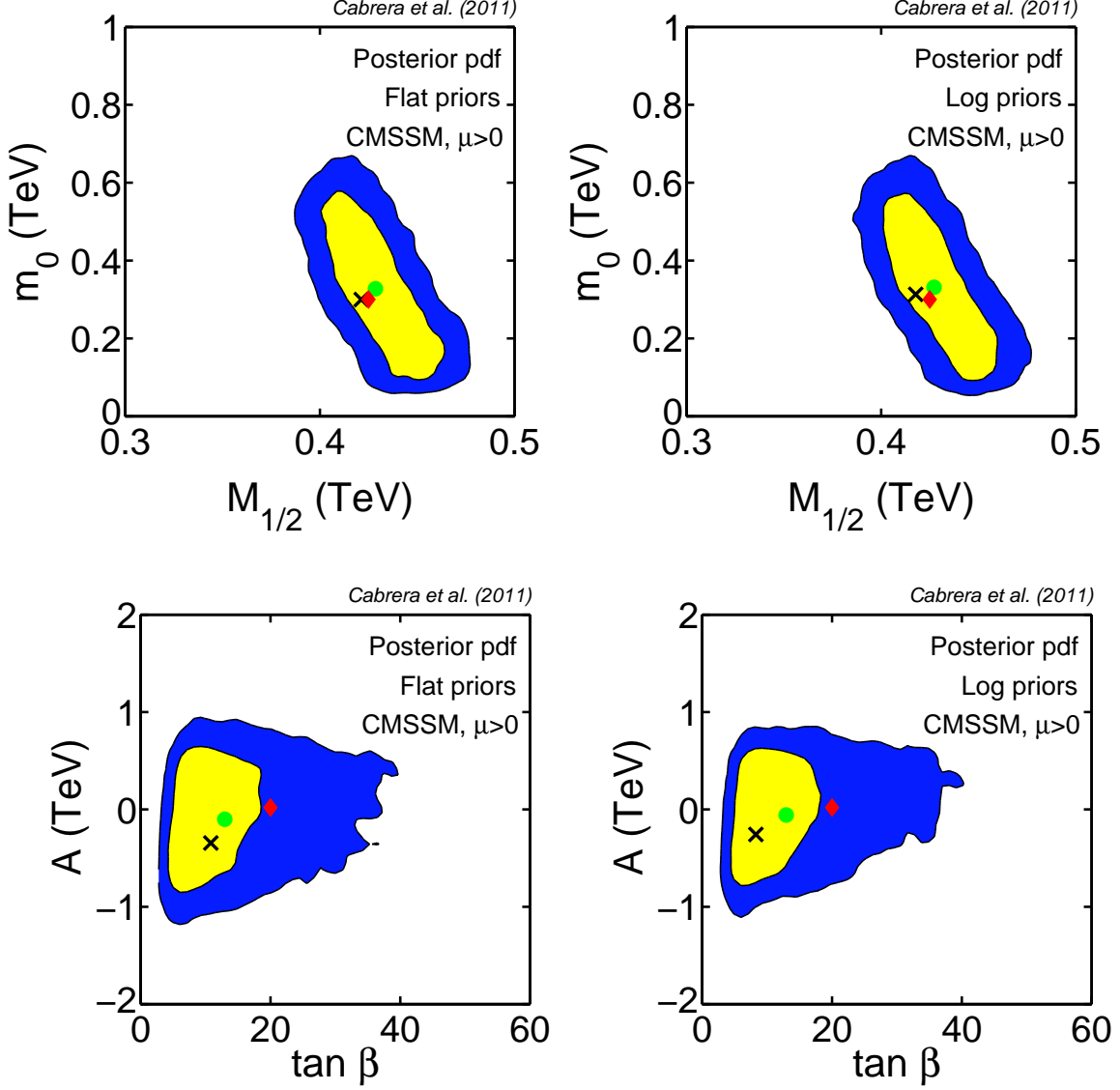


Figure 3: As Fig. 2 but using both normalization and shape. Note the different ranges of the two figures.

the parameter space we use the modified formula (6.4):

$$\begin{aligned} \text{LHC - likelihood} &\propto \mathcal{P}(f = \frac{v}{Lu}) \\ &\times \prod_{i=1}^K \left(\frac{(u_i + v_i)!}{u_i! v_i!} \int dg_i \frac{1}{g_i} \left(\frac{v}{u} g_i \right)^{v_i} \left(1 + \frac{v}{u} g_i \right)^{-1-u_i-v_i} \mathcal{P}(g_i) \right) \end{aligned} \quad (4.10)$$

We recall that $\mathcal{P}(g)$ carries all the systematic uncertainties affecting the shape of the histograms, and is given by the gaussian (3.11) with $\Delta_g = 0.2$; as explained in the previous

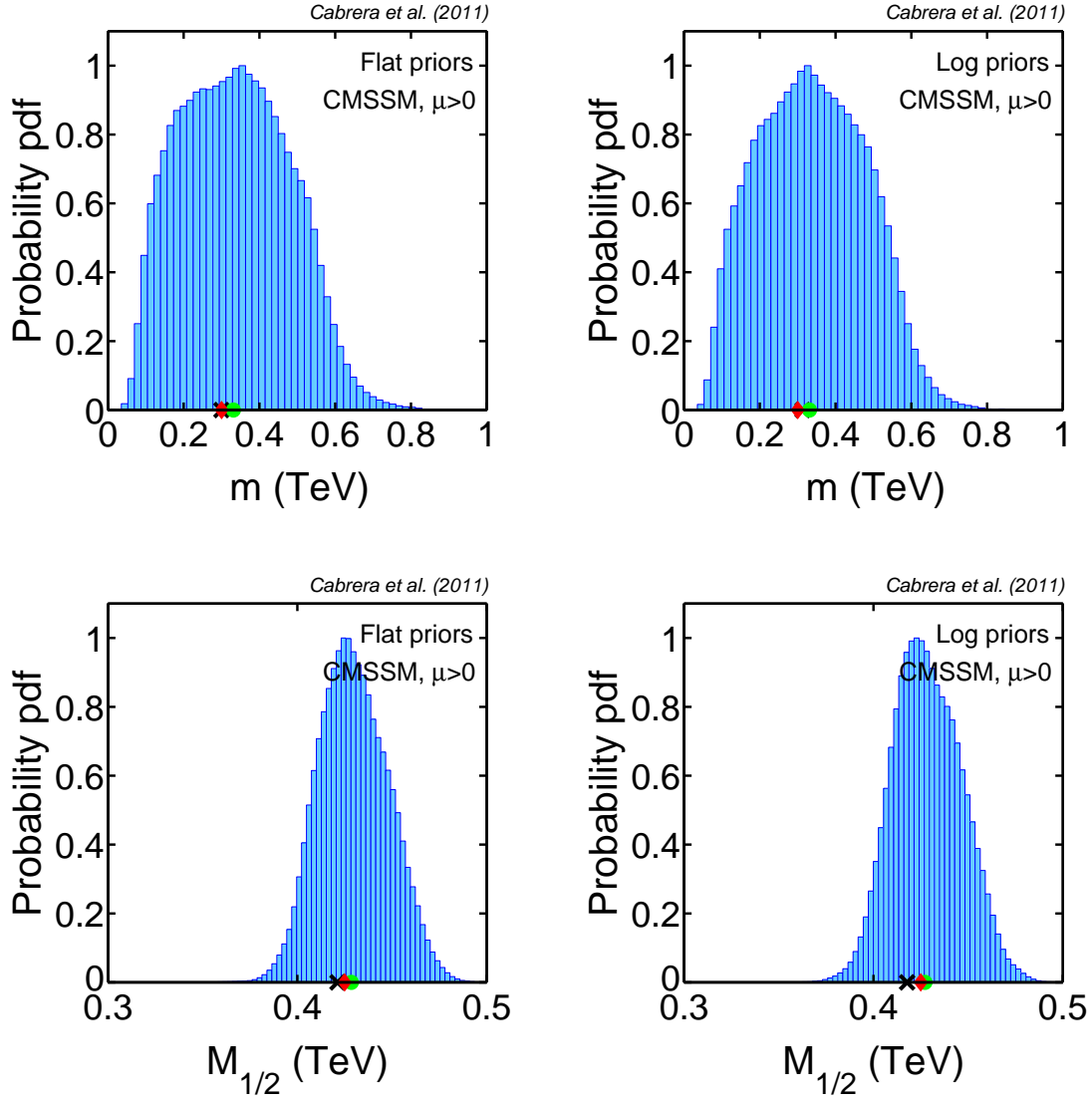


Figure 4: 1D marginalized posterior probability density function of the m and $M_{1/2}$ parameters (upper and lower panels respectively) for flat (left panels) and logarithmic (right panels) priors. The small filled circle represents the mean value of the posterior pdf, the cross corresponds to the best-fit point and the diamond to the SU9 model.

subsection. Note that, as could be expected, the correction due to the difference in luminosity does not affect the shape-part of the likelihood.

Fig. 3 is as Fig. 2, but after including the likelihood associated to the shape in the analysis. The upper panels show, for log and flat priors, the posterior pdf in the $M_{1/2} - m$ plane, after marginalizing the rest of the parameters. As expected, the test of the theory is now much more efficient and the previous degeneracies disappear (note the different ranges of the two figures). This illustrates the potential of making use of all the information contained in the theoretical and experimental histograms when computing the likelihood of a model, provided the various sources of uncertainty are properly taken into account.

The lower panels of Fig. 3 show the posterior in the $\tan\beta - A$ plane. Again, the cross section of the type of events considered does not depend appreciably on these parameters, which is reflected in the plots. Still, introducing the test for the shape slightly improves the sensitivity of the search to the values of A and $\tan\beta$, but that sensitivity is anyway very small.

Figs. 4 and 5 show, for logarithmic and flat priors, the unidimensional posteriors for m , $M_{1/2}$, A and $\tan\beta$, after marginalization of all the parameters, except the one plotted in each graph. The shape of these functions reflects the previous discussion. It is worth-noticing the great precision in the determination of the gaugino mass, which comes from the fact that, due to the renormalization group running, $M_{1/2}$ is the parameter that dominantly determines the low-energy spectrum of the CMSSM.

Finally, we note that the posteriors have in all cases a very slightly dependence on the type of prior used, reflecting the robustness of the approach.

5. Conclusions

Due to the complexity of the LHC experiment, much of the comparison between LHC data and theoretical predictions has to be made by confronting experimental histograms (in different variables) and theoretical histograms produced by simulations. In many cases the comparison is performed by comparing the total number of events after choosing a clever variable and applying convenient cuts. Other techniques make use of particular features of the histograms, like the presence of endpoints. But the procedure can be optimized by evaluating the actual likelihood associated to the complete histogram. The main goal of this paper has been precisely to present a rigorous and effective method to compare experimental and theoretical histograms, evaluating the total likelihood, and apply it to a physically relevant case. In doing this we have taken into account that, besides the statistical uncertainties inherent to the histograms, there are additional sources of systematic error.

In the method presented, the complete likelihood is rigorously separated into two factors: the likelihood of the total number of events and the likelihood of the shape of the histogram. This in turn allows to treat the corresponding sources of systematic uncertainty in a separate way as well. This is very convenient when there are reasons to expect different systematic errors in the two pieces. The final formula for the total likelihood is given in eq.(3.17).

The procedure can be easily incorporated to both frequentist and Bayesian analyses, since both are based on the likelihood of the theoretical models. In the two approaches, incorporating the total likelihood optimizes the chances of picking up a signal of new physics and, once the signal is found, identifying which new physics is behind. E.g. if the

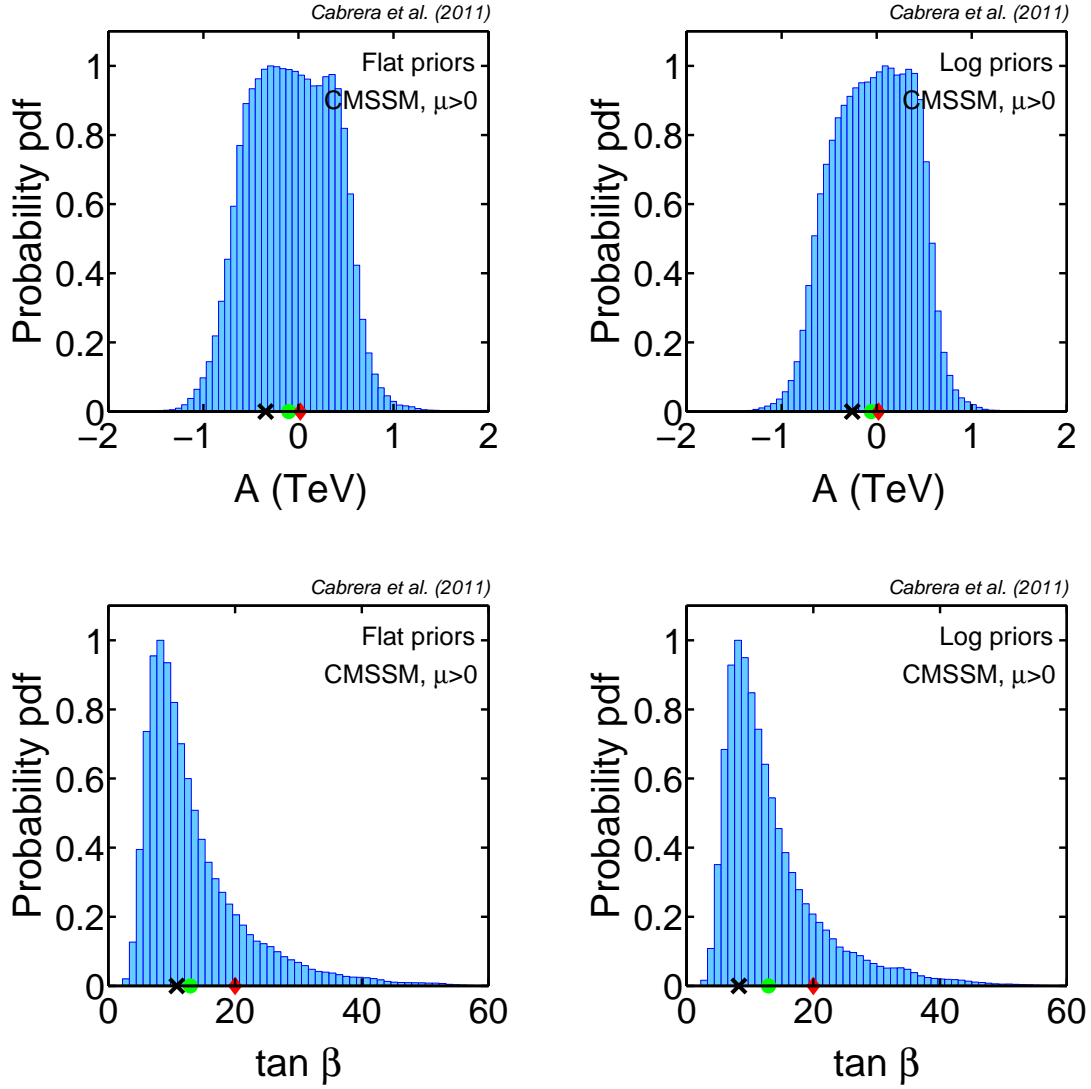


Figure 5: As Fig. 4 but for A and $\tan \beta$ parameters.

new physics is supersymmetry, it allows to find in an optimal way the parameters of the supersymmetric model.

We have illustrated the latter point by showing how a search in the CMSSM parameter space, using Bayesian techniques, can effectively find the correct values of the CMSSM parameters by comparing histograms of events with multijets + missing transverse momentum displayed in the effective-mass variable. The procedure is in fact very efficient to identify the true supersymmetric model, in the case supersymmetry is really there and accessible to the LHC. But, of course, the technique can be applied to any scenario of new physics.

6. Appendix A: Histogram comparison when experiment and the simulation have different luminosities

Some expressions of sects. 2 and 3 have to be modified when the effective luminosity of the simulation is not the same as the experimental one. In practice, the former can change from point to point when scanning the parameter space since typically one simulates a fixed number of supersymmetric events (say 10^4 events), but obviously the cross section changes throughout the parameter space. Of course one could adjust at every point the luminosity so that it coincides with the experiment, but normally this is costly in running time, and it is unnecessary, since the comparison can still be made as described next.

Let us call L^{th} , L^{exp} the luminosities of the theoretical simulation and the experiment, respectively, and suppose for a moment there are no systematic errors. Then the means that, under the null-hypothesis, are responsible for the experimental data, v_i , are *not* the ones of the simulation, say $\hat{\mu}_i$, but

$$\mu_i = \frac{L^{exp}}{L^{th}} \hat{\mu}_i \equiv L \hat{\mu}_i . \quad (6.1)$$

Hence eq.(2.8) becomes

$$P(v_i|u_i) = \int \prod_{i=1}^K d\hat{\mu}_i \frac{(L\hat{\mu}_i)^{v_i}}{v_i!} e^{-L\hat{\mu}_i} \frac{\hat{\mu}_i^{u_i}}{u_i!} e^{-\hat{\mu}_i} = \prod_{i=1}^K \frac{(u_i + v_i)!}{u_i! v_i!} L^{v_i} (1 + L)^{-1-u_i-v_i} . \quad (6.2)$$

Once systematic uncertainty is taken into account, see section 3, everything is actually easy to handle since the luminosity factor L plays the role of a systematic and universal factor affecting the means of the simulation. More precisely, the equation (3.4), that relates the true means to be compared with the experiment, μ_i , with those of the simulation, μ_i^{th} , becomes

$$\mu_i = L f g_i \mu_i^{th} . \quad (6.3)$$

Therefore the subsequent equations remain the same with the simple change $f \rightarrow Lf$. In particular, the likelihood given by eq.(3.17) becomes now

$$\mathcal{P}(v_i|u_i) \propto \mathcal{P}(f = \frac{v}{Lu}) \prod_{i=1}^K \left(\frac{(u_i + v_i)!}{u_i! v_i!} \int dg_i \frac{1}{g_i} \left(\frac{v}{u} g_i \right)^{v_i} \left(1 + \frac{v}{u} g_i \right)^{-1-u_i-v_i} \mathcal{P}(g_i) \right) \quad (6.4)$$

This is the formula we have used in our scan of the CMSSM parameter space.

Acknowledgements

We thank F. Feroz and R. Trotta for interesting discussions and suggestions.

This work has been partially supported by the MICINN, Spain, under contract FPA 2007–60252, the Comunidad de Madrid through Proyecto HEPHACOS S-0505/ESP–0346, and

by the European Union through the UniverseNet (MRTN-CT-2006-035863). M. E. Cabrera acknowledges the financial support of the CSIC through a predoctoral research grant (JAEPRe 07 00020). The work of R. Ruiz de Austri has been supported in part by the project PARSIFAL (FPA2007-60323) of the Ministerio de Educación y Ciencia of Spain and by the Spanish MICINN's Consolider-Ingenio 2010 Programme under the grant MULTIDARK CSD2009-00064, CPAN CSD2007-00042 and PAU CSD2007-00060. V.A.M. acknowledges support by the Spanish Ministry of Science and Innovation (MICINN) under the project FPA2009-13234-C04-01, by the Ramón y Cajal contract RYC-2007-00631 of MICINN and CSIC, by the Spanish Agency of International Cooperation for Development under the PCI project A/030322/10 and by the grant UV-INV-EPDI11-42955 of the University of Valencia. The use of the Hydra cluster of the IFT-UAM/CSIC is also acknowledged.

References

- [1] G. L. Bayatian *et al.* [CMS Collaboration], J. Phys. G **G34** (2007) 995-1579.
- [2] G. Aad *et al.* [The ATLAS Collaboration], arXiv:0901.0512 [hep-ex].
- [3] CMS Collaboration, CMS Physics Analysis Summary SUS-09-001 (2009).
- [4] <https://twiki.cern.ch/twiki/bin/view/AtlasPublic>.
- [5] <https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResults>.
- [6] G. L. Kane, C. F. Kolda, L. Roszkowski and J. D. Wells, Phys. Rev. D **49** (1994) 6173 [arXiv:hep-ph/9312272].
- [7] R. Trotta, *Contemporary Physics*, **49**, 2, (2008) 71-104.
- [8] For a review see S. P. Martin, arXiv:hep-ph/9709356.
- [9] J. O. Berger, B. Liseo and R. L. Wolpert, Statistical Science 1999, Vol. 14, No. 1, 1-28
- [10] M. E. Cabrera, J. A. Casas and R. Ruiz de Austri, JHEP **0903**, 075 (2009) [arXiv:0812.0536 [hep-ph]].
- [11] M. E. Cabrera, J. A. Casas and R. Ruiz d Austri, JHEP **1005**, 043 (2010) [arXiv:0911.4686 [hep-ph]].
- [12] B. C. Allanach, Comput. Phys. Commun. **143** (2002) 305 [arXiv:hep-ph/0104145].
- [13] S. P. Martin, In *Kane, G.L. (ed.): Perspectives on supersymmetry* 1-98. [arXiv:hep-ph/9709356 [hep-ph]].
- [14] L. E. Ibanez, C. Lopez, Nucl. Phys. **B233** (1984) 511.
- [15] R. R. de Austri, R. Trotta and L. Roszkowski, JHEP **0605** (2006) 002 [arXiv:hep-ph/0602028].
- [16] ATLAS Collaboration, Results presented at EPS 2011.
- [17] CMS Collaboration, CMS-PAS-SUS-11-004 (2011).
- [18] T. Sjostrand, S. Mrenna, P. Z. Skands, JHEP **0605** (2006) 026. [hep-ph/0603175].

- [19] Available from: <http://superbayes.org/>.
- [20] F. Feroz and M. P. Hobson, *Mon. Not. Roy. Astron. Soc.* **384** (2008) 449–463.
- [21] F. Feroz, M. P. Hobson, and M. Bridges, *Mon. Not. Roy. Astron. Soc.* **398** (2009) 1601–1614.
- [22] M. E. Cabrera, J. A. Casas, A. Delgado, [arXiv:1108.3867 [hep-ph]].
- [23] H. Baer, C. h. Chen, M. Drees, F. Paige and X. Tata, *Phys. Rev. D* **59**, 055014 (1999) [arXiv:hep-ph/9809223].